A close up of a sign

Description automatically generated with low confidence

**Department of Computer Applications**

**MCA – A SECTION (Batch- 2021-2023)**

**Session: May 2023 to August 2023**

**PROJECT WORK – SYNOPSIS**

**TITLE OF THE PROJECT : DiagnoseMe : A Predictive Health Analysis for Early Disease Detection**

**STUDENT NAME & SECTION : Nakul S Anand T - MCA-A**

**REGISTRATION NO. : 21P01068**

**GUIDE NAME : Mr. N Karthik**

**PROJECT STATUS (8TH OF JUNE):**

* I acquired relevant datasets from the Machine Learning Repository and Kaggle. These datasets are essential for my project and provide the necessary information for analysis and modeling.
* I installed the required packages and libraries in Python to support my project. These packages include data manipulation, statistical analysis, visualization, and machine learning tools.
* I established a connection to SQLite3 and created a new database named "DiagnoseMe." Within this database, I created tables to store and organize the data from the acquired datasets.
* Module 1 completed.
* I started analyzing the first dataset, which focuses on diabetes. This involved performing statistical analysis to gain a better understanding of the data and its characteristics.
* I checked the dataset for missing values. Finding and understanding missing data ensures the integrity and completeness of the dataset for further analysis.
* Utilizing boxplots, I identified potential outliers in the dataset. Outliers are data points that deviate significantly from the rest and can have a notable impact on analysis and modeling results.
* To aid in the analysis process, I created visualizations, such as histograms, to explore the distribution of variables in the dataset. Visualizations provide a clear and intuitive representation of the data and assist in identifying patterns and trends.
* I conducted correlation analysis to examine the relationships between variables in the dataset. This involved calculating correlation coefficients and visually representing them using a heatmap. Correlation analysis helps to understand the interdependencies among variables.
* I employed the Shapiro-Wilk test to assess the normality of the data distribution. Normality testing is essential for many statistical analyses and provides insights into the appropriateness of certain modeling techniques.
* Started basic analysis on dataset 2.

Overall, the project is progressing well, and these initial steps have laid a solid foundation for further analysis and modeling. The acquired datasets, database setup, statistical analysis, visualization, and normality testing are vital components of my project, allowing me to gain valuable insights into the data.

In the upcoming week I’ll be repeating the whole Exploratory Data analysis (Module -2) for rest of the datasets.